DOCUMENT TYPE DEFINITION GENERATING METHOD AND APPARATUS, AND STORAGE MEDIUM FOR STORING PROGRAM

BACKGROUND OF THE INVENTION

Field of the Invention

5   The present invention relates to a computerized document processing executed by a personal computer, a word processor, and the like, particularly to a method and apparatus for generating the document type

10   definition of a structured document, and a storage medium in which a program is stored.

Related Background Art

In recent years, the computerized documents prepared by a personal computer, a word processor, and

15   the like have widely been used. The introduction of a structured document is advanced in which the computerized document is consistently treated and the elements constituting the document are provided with semantic information. In this structured document,

20   each document element is held between front and back tags including element names (tag names), and in many cases description is performed for each document type in accordance with the document type definition of defining a place, order, frequency and the like in

25   which the element appears.

On the other hand, the structured document can be described without preparing the document type

definition. However, when the documents prepared by a plurality of users are integrated to form one document, and if the individual users use the tags having arbitrary titles, there is a possibility of attaching

5   different tag names to the same element, or conversely attaching the same tag name to different elements.

In this case, there arise problems that the semantic information attached to the tag cannot correctly be handled, and that redundancy is generated

10   with respect to the tag.

SUMMARY OF THE INVENTION

An objective of the present invention is to provide a method and apparatus for generating document

15   type definition from a structured document provided with tags, and a storage medium which stores the program.

Another objective of the present invention is to provide a document type definition generating method

20   and apparatus which can correctly treat semantic information given to tags, and a storage medium which stores the program.

Further objective of the present invention is to provide a document type definition generating method

25   and apparatus which can generate document type definition with redundancy to tags removed therefrom, and a storage medium which stores the program.

According to one aspect, the present invention which achieves these objectives relates to a document processing method comprising: in a structured document provided with a tag having an element name in each document element, a physical structure judging step of judging a physical structure of each document element; a semantic structure judging step of judging a semantic structure of the document element; and a document type definition generating step of generating document type definition to define appearance state of the document element in the structured document based on judgment results of the physical structure judging step and the semantic structure judging step.

According to another aspect, the present invention which achieves these objectives relates to a document processing apparatus comprising: in a structured document provided with a tag having an element name in each document element, physical structure judging means for judging a physical structure of each document element; semantic structure judging means for judging a semantic structure of the document element; and document type definition generating means for generating document type definition to define appearance state of the document element in the structured document based on judgment results of the physical structure judging means and the semantic structure judging means.

According to still another aspect, the present invention which achieves these objectives relates to a computer-readable storage medium storing a document type definition generating program for controlling a computer to perform document type definition generation, the program comprising codes for causing the computer to perform, in a structured document provided with a tag having an element name in each document element, a physical structure judging step of judging a physical structure of each document element, a semantic structure judging step of judging a semantic structure of the document element, and a document type definition generating step of generating document type definition to define appearance state of the document element in the structured document based on judgment results of the physical structure judging step and the semantic structure judging step.

Other objectives and advantages besides those discussed above shall be apparent to those skilled in the art from the description of a preferred embodiment of the invention which follows. In the description, reference is made to accompanying drawings, which form a part thereof, and which illustrate an example of the invention. Such example, however, is not exhaustive of the various embodiments of the invention, and therefore reference is made to the claims which follow the description for determining the scope of the invention.

BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 is a block diagram of a document type definition generating apparatus.

Fig. 2 is a flowchart showing the procedure of a document type definition generation processing.

Figs. 3A and 3B are diagrams showing examples of structured document data.

Fig. 4 is a flowchart showing the processing procedure of physical structure analysis.

Fig. 5 is a flowchart showing the processing procedure of semantic structure analysis.

Fig. 6 is a flowchart showing the processing procedure of removing tag redundancy.

Fig. 7 is a diagram showing one example of document type definition.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

A preferred embodiment of the present invention will be described hereinafter with reference to the accompanying drawings.

<First Embodiment>

Fig. 1 is a block diagram of a document type definition generating apparatus according to the present invention.

In Fig. 1, an input unit 101 is constituted of a keyboard, a pointing apparatus, and the like, and is used for a user to input data or commands. An external

memory unit 102 is constituted of a storage apparatus using media such as a hard disk to store structured document data as a processing object, data of semantic information database (DB) described later, generated document type definition, and the like. A display unit 103 is constituted of CRT, a liquid crystal display, and the like to display the structured document data, the generated document type definition, and the like.

A CPU 104 performs control of each component of the apparatus, reads and executes a program, and realizes various processings. A ROM 105 stores fixed data and program. A control program for realizing a processing procedure as described later with reference to the flowcharts of Fig. 2 to 6 may be stored in the ROM 105, or read from the external memory unit 102. A RAM 106 presents an operation area necessary for the processing of the apparatus. A bus 107 connects the apparatus components.

Fig. 2 is a flowchart showing the procedure of a document type definition generation processing according to the present invention.

First, the structured document is inputted in step S201. This is executed by reading the structured document from the external memory unit 102. One example of the structured document given herein is shown in Fig. 3A. For example, a first line "<Title>" indicates a start tag, "</Title>" indicates an end tag,

and "TV SET OPERATING INSTRUCTIONS" held between these tags is a document element indicating a tag content. Moreover, "Title" is an element name (tag name). Furthermore, the attribute and value of the element can
5   be described in the tag.

In the next step S202, each tag position is detected from the structured document, and a tag number is attached in order from the top "<Title>".

Subsequently, in step S203, the physical structure
10   in the document is detected. For example, in Fig. 3B, as diagrammatically represented in "<Para>" indicating a paragraph, a feature that a sentence group starting with an indention is regarded as the paragraph is detected. The processing procedure for detecting such
15   physical structure is shown in the flowchart of Fig. 4.

First, in step S401, a line in which indention is performed is found in the document, and in the next step S402 the sentence group following the line is detected. In this case, the line in which the
20   indention is performed to the line in which the next indention is performed, or to the line right before a blank line can be set to the sentence group. In this case, the indention (double indention) performed in quotation in which the quotation is represented by
25   performing the indention, and blank lines described by constantly skipping one or more lines are excluded as structures meaningless for the detection of the

physical structure from the entire document pattern to perform the processing in the step S402.

Turning back to Fig. 2, in the next step S204, the semantic structure of the inputted structured document is detected. As one example, in Fig. 3A, the contents of tags "<Section>" have forms in which "1.", "2.", "3." are attached to top positions. Here, the content of tag "<Section>" can semantically be presumed to have "numeral." on its top. One example of processing procedure for detecting the semantic structure is shown in the flowchart of Fig. 5.

First, in step S501, communication is performed with a semantic information database (DB) 51 with respect to all words and codes in the document to provide the connection between words in the document and the types of words and codes. In the next step S502, the semantic structure found in each document element is detected based on this result.

Returning to Fig. 2, in the next step S205, a first appearing tag is regarded as the tag to be processed, and it is judged in step S206 whether or not the processing of the tag is all completed.

When the tag processing is not completed, the process shifts to step S207, in which the tag as the present processing object, and the information on the physical and semantic structures detected in the steps S203 and 204 are unified. Here, the unifying means

that when physical and semantic features are present in the line related with the tag used as the present processing object, the tag and the information are connected. Subsequently, in step S208, the process is moved to the next appearing tag, thereby returning to the step S206.

On the other hand, when it is judged in the step S206 that the tag processing is all completed, the process shifts to step S209, in which similarity is obtained between the tags having different titles. When the similarity is equal to or more than a predetermined threshold value, the tags are regarded as the same tag, and one of the tags is prevented from appearing on the document type definition to be generated. The processing procedure for obtaining this similarity to determine whether or not the tags have the same content is shown in the flowchart of Fig. 6.

First, the similarity of tags A, B having different titles is calculated in step S601. This calculating method comprises setting the similarity of the physical structure to 1 when the physical structures agree with each other. When the physical structures do not completely agree with each other, but partially agree with each other, the similarity of the physical structure is set to a value less than 1 which corresponds to the agreed proportion. The similar concept is applied to the semantic structure, and the

similarity of the semantic structure is obtained. The
dividing of the sum of the similarity of the physical
structure and the similarity of the semantic structure
by 2 results in a general similarity d_{AB} of A and B.

5   In the next step S602, the similarity d_{AB} obtained
in the step S601 is compared with the predetermined
threshold value δ. When the similarity d_{AB} is less than
δ, the process jumps to step S604 for trial of the next
combination.

10   When the similarity d_{AB} is equal to or more than
the threshold value δ, the process shifts to step S603,
in which the tag B is regarded as being of the same
type as the tag A, the tag B is finally struck off a
list for generating the document type definition, and
15 redundancy is removed.

When the processing of the step S603 is completed,
the process advances to step S604, in which it is
judged whether or not the trial of combination of all
tags is made. When the combination of all tags is not
20 tried, the process returns to the step S601. When the
combination of all tags is tried, the subroutine
processing is ended to return to the main routine of
Fig. 2.

Moreover, in the step S209, in addition to the
25 above-described processing of Fig. 6, the physical
structure and semantic structure of the document
elements having the same title are compared. When the

structures are different, the title of one of the tags is changed. For this purpose, the similarity is obtained between tags Aa and Ab having the same tag name in the same manner as described above. When

5   similarity value d_{AaAb} is less than the threshold value, the title of the tag Ab is changed. This threshold value may be different from the above-described value.

In step S210 of Fig. 2, the sentence word between the start tag and the end tag which have the same title

10 is analyzed to obtain the information to be included in the tags. This analysis result is used to generate the document type definition in the next step S211.

Fig. 7 is a diagram showing one example of the generated document type definition, and the document

15 type definition generated from the structured document data shown in Fig. 3A is shown as document type "manual".

Here, in Fig. 3A, the content of tag <Sect> agrees in physical structure with the content of tag

20 <Section>, and the tags are the same in semantic structure in that they have the form of "numeral.". Therefore, it is determined in the step S209 that the tag <Sect> has the same content as that of the tag <Section>. As a result, the generated document type

25 definition does not use <Sect>, and in <Body>, Section+, that is, tag <Section> repeatedly appears.
<Second Embodiment>

In the above-described first embodiment, the physical and semantic structures in the document are judged based on the sentence (portions other than tags), but the present invention is not limited to

5   this.

For example, the physical information such as the relative positional relation between the tags and the inclusive relation of the tags is detected as the physical structure, or the meaning represented by the

10   tag name or attribute is detected as the semantic structure, so that these structures may be used as the objects to obtain the similarity.

According to the embodiments described above, since the physical and semantic structures of the

15   document element surrounded with the tags are judged, and the document type definition of the structured document provided with the tags is generated, the semantic information given to the tags can correctly be treated.

20   Furthermore, the redundancy to the tags having the same content can be removed, and the document type definition can be generated in which there are no tags being the same in title and different in meaning.

Additionally, the present invention may be applied

25   to a computer system constituted of a plurality of apparatuses (e.g., host computer, interface apparatus, reader, printer, and the like), or to a device

constituted of one apparatus (e.g., word processor, copying machine, facsimile device, and the like).

Moreover, it goes without saying that the objective of the present invention can be achieved by supplying a storage medium storing the program code of software to realize the function of the above-described embodiment to the system or the device, and reading and executing the program code stored in the storage medium by the computer (or CPU or MPU) of the system or the device.

In this case, the program code itself read from the storage medium realizes the function of the above-described embodiment, and the storage medium in which the program code is recorded constitutes the present invention.

As the storage medium in which the program code, and tables and other variable data are stored, for example, a floppy disk (FD), a hard disk, an optical disk, an optomagnetic disk, CD-ROM, CD-R, a magnetic tape, a nonvolatile memory card (IC memory card), ROM, and the like can be used.

Moreover, the function of the above-described embodiment is realized by executing the program code read by the computer, but it goes without saying that the present invention also includes a case in which an operating system (OS) operating on the computer performs a part or the whole of an actual processing

based on the instruction of the program code and the
function of the above-described embodiment is realized
by the processing.

Although the present invention has been described
5   in its preferred from with a certain degree of
particularity, many apparently widely different
embodiments of the invention can be made without
departing from the spirit and the scope thereof. It is
to be understood that the invention is not limited to
10   the specific embodiments thereof except as defined in
the appended claims.